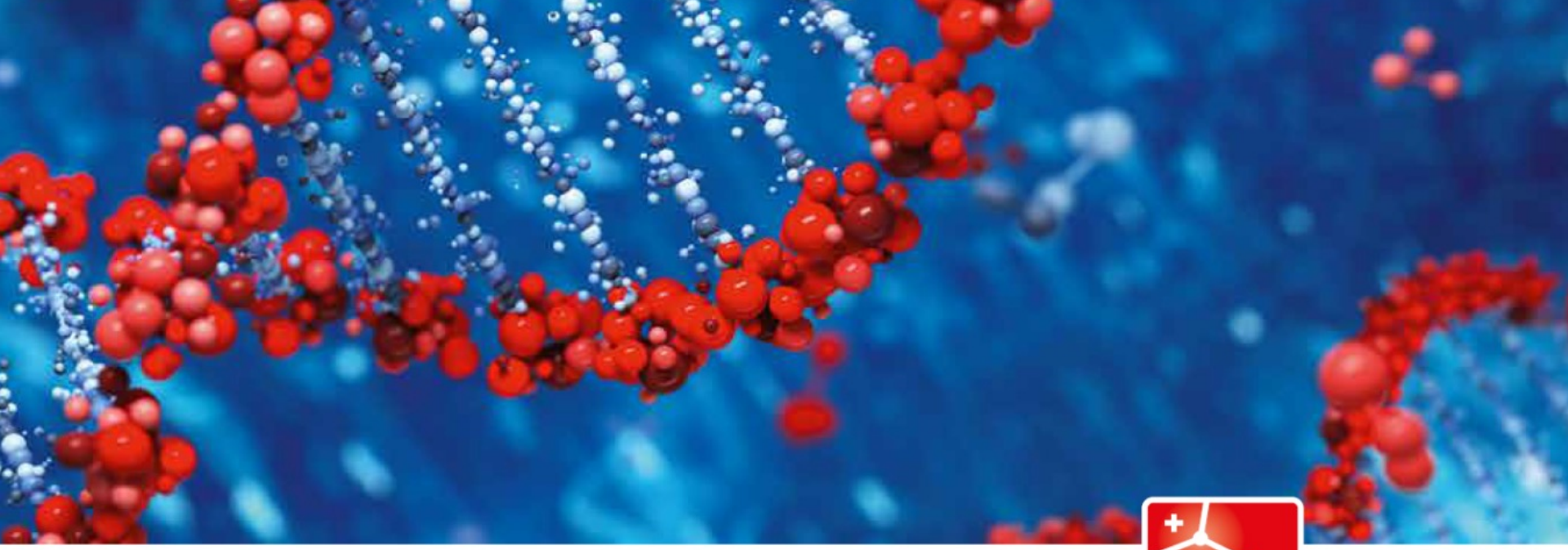


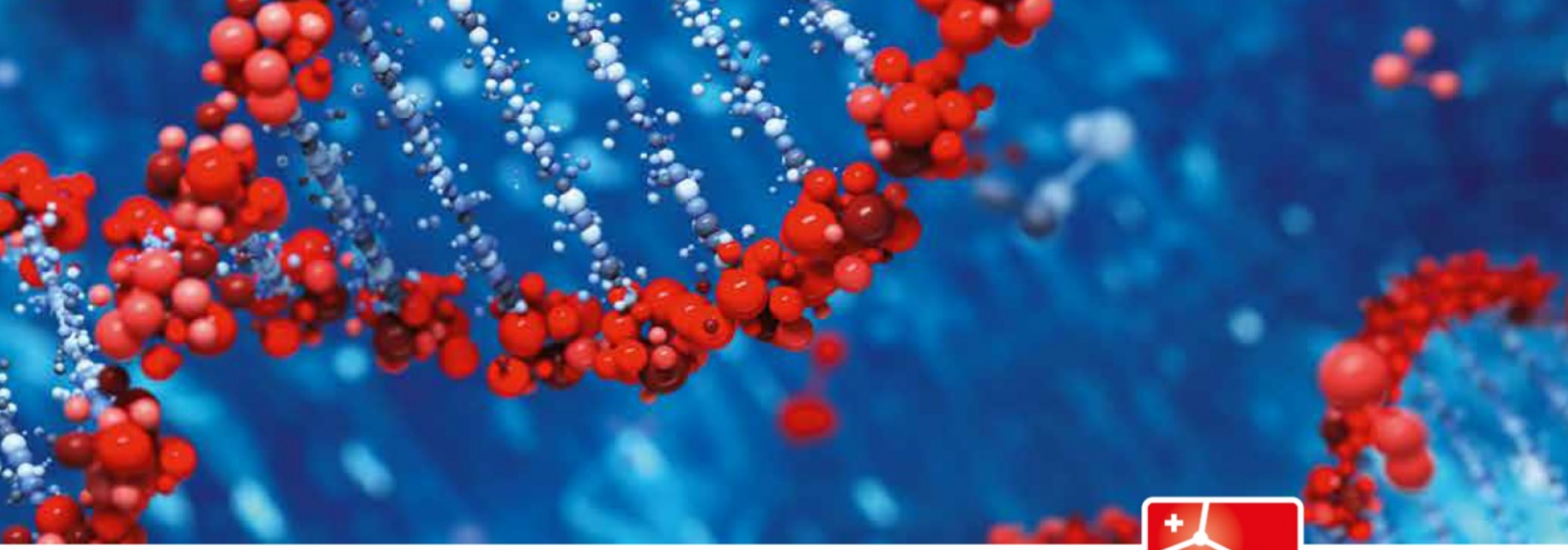
Swiss Institute of Bioinformatics

# Overview of bioinformatics analyses for metagenomics data

*Aitana Lebrand*
*SIB Swiss Institute of Bioinformatics – Clinical Bioinformatics*
*ICCMg, Geneva, 13th October 2016*

elixir SWITZERLAND

www.sib.swiss



Swiss Institute of Bioinformatics

# Overview of bioinformatics analyses for metagenomics data *for non-bioinformaticians*

*Aitana Lebrand*
*SIB Swiss Institute of Bioinformatics – Clinical Bioinformatics*
*ICCMg, Geneva, 13th October 2016*

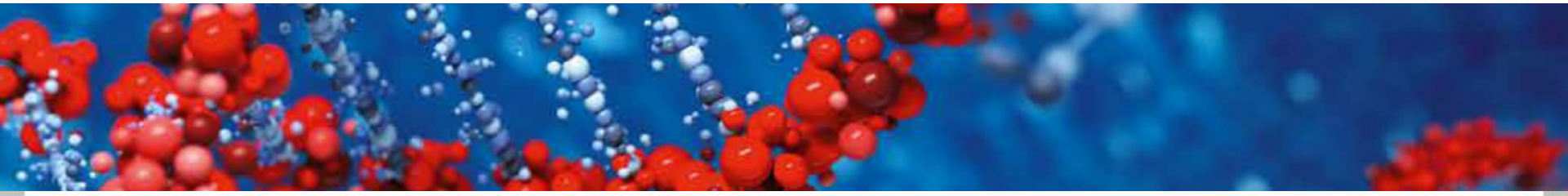
elixir SWITZERLAND

www.sib.swiss

I. Analysis pipelines for metagenomics data

II. Harmonizing best practices in clinical metagenomics across Switzerland

# I. Analysis pipelines for metagenomics data



*!! Apologies if I do not mention your favourite tool !!*

# Clinical metagenomics questions

Patient sample

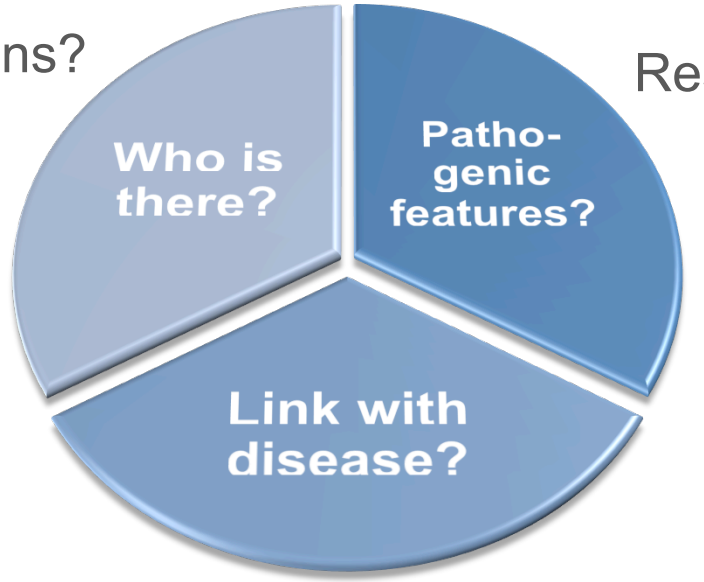
One analysis

What pathogens?

What bacteria?

What virus?

What fungi?




Resistance?

Virulence?

Pathways?

Transmission routes?

Markers of disease?

# Clinical metagenomics for diagnosis

consultation → sample → sequencing → pre-processing → analysis → report → consultation






---------- Bioinformatics ----------

Data management

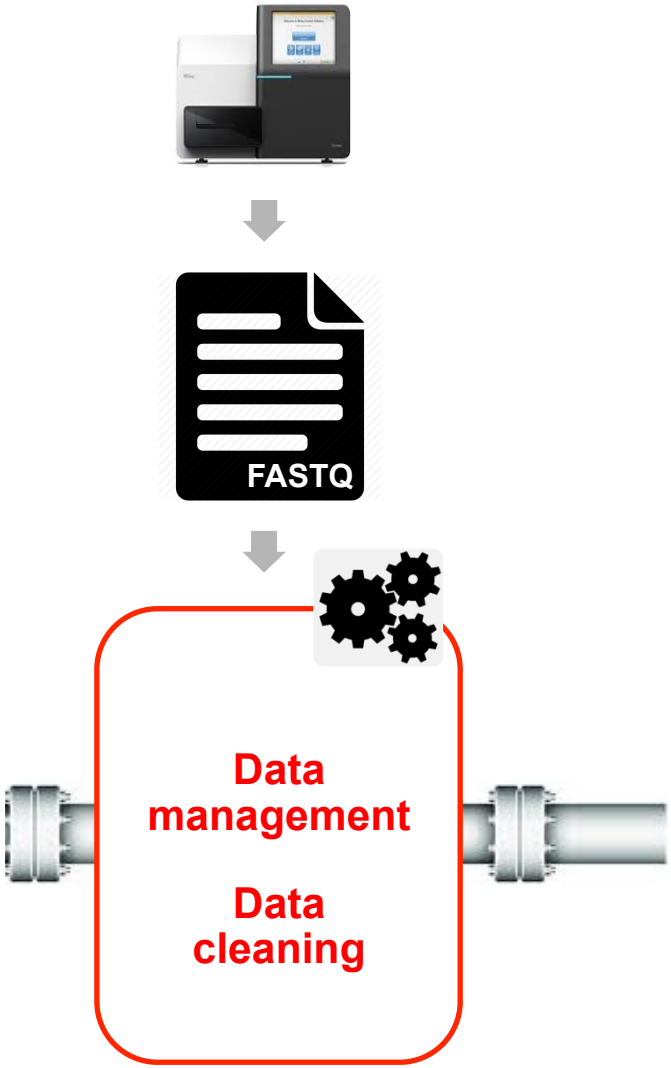
Data cleaning

Analysis and interpretation

Integration of information from specialized databases


Consolidated reports to support diagnosis, expert view and treatment decisions

# Pre-processing




*“Get clean data for downstream and future analyses”*

# NGS data pre-processing – trimming & filtering

Sequence →
Quality scores →

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTAT
+
BBBFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIFFFIII
```

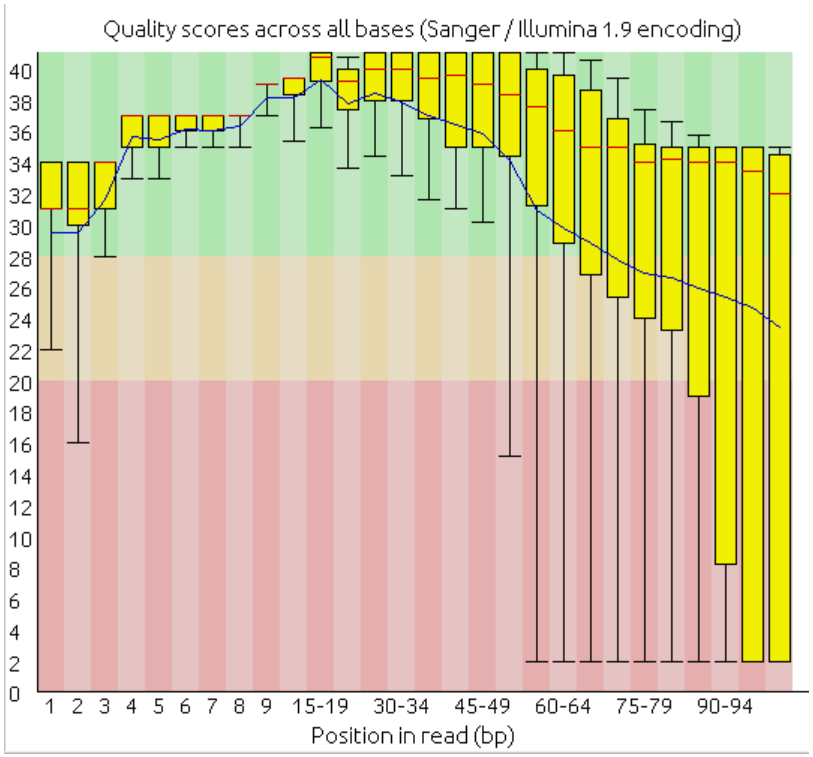
FASTQ

**Trimming**: clip only certain regions of the read

✓ **Get rid of adaptor and low quality regions**

**Filtering**: remove reads that do not meet quality criteria

✓ **Get rid of contaminants, duplicates**



# Analysis

*"Taxonomic Profiling"*


*"Functional analyses"*

*"Comparative metagenomics"*

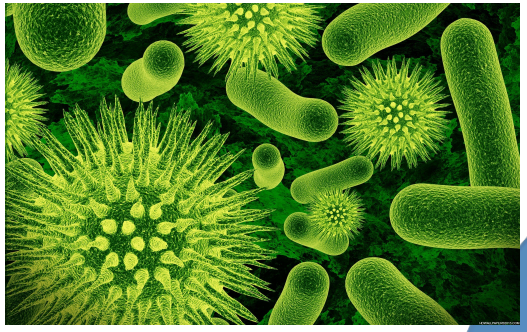
Data management

Data cleaning

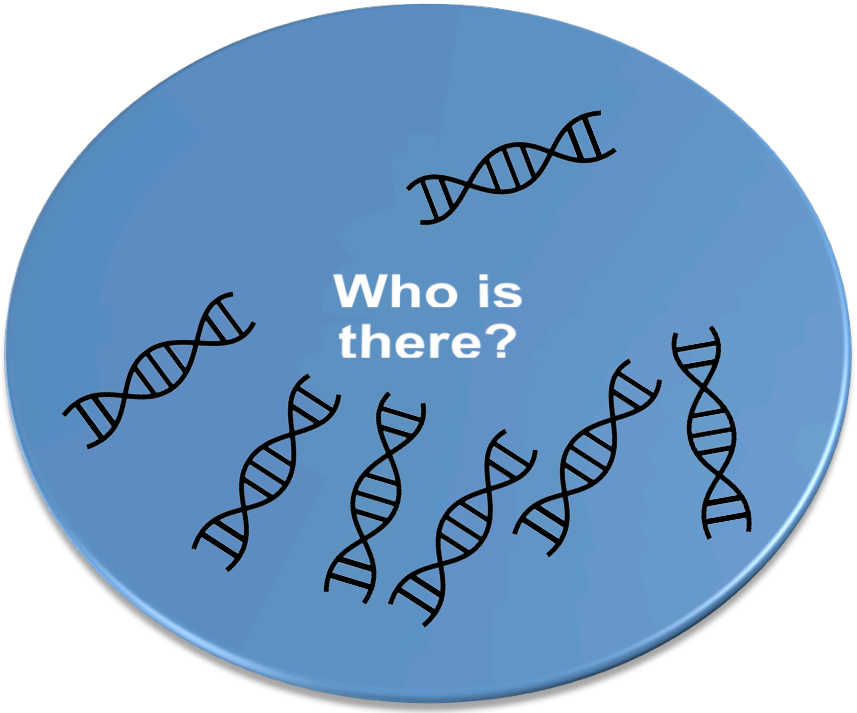
**Analysis and interpretation**

**Integration of information from specialized databases**

# Taxonomic profiling




# Taxonomic profiling



# Taxonomic profiling – marker based

**Sequence/Label only marker genes**

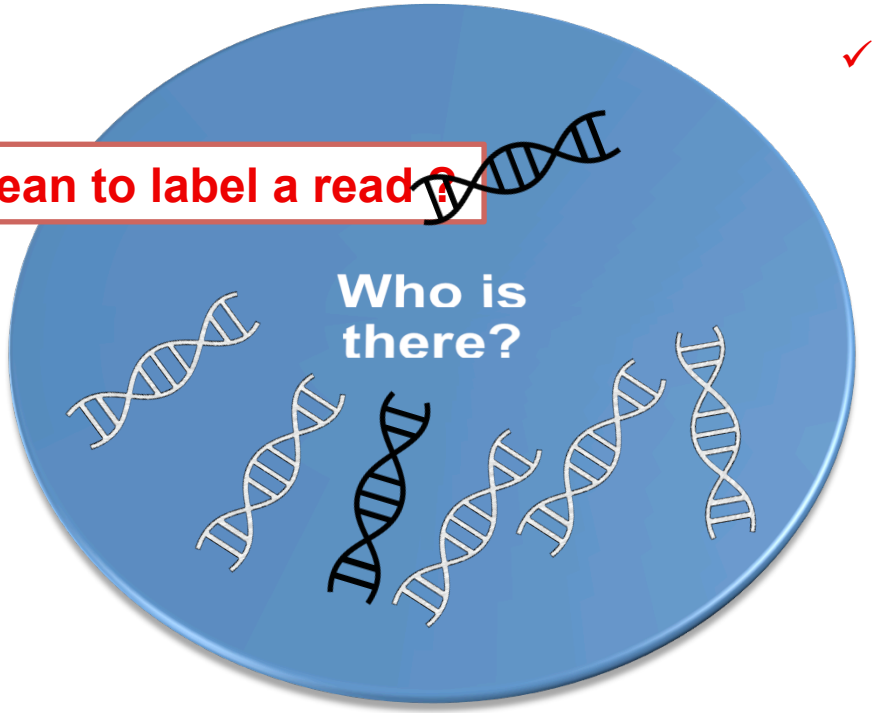

- ✓ Universal marker (16S, ITS…)

# Taxonomic profiling – marker based

## Sequence/Label only marker genes

- ✓ Universal marker (16S, ITS…)

**What does it mean to label a read?**




**GG, SILVA, RDP, UNITE…**
**→ OTU**

# Sequence alignment in a nutshell



! **Short reads** are more likely to occur by chance in the database
→ may not be significant.

**BLAST**
RefSeq, nr/nt, nr, etc.

**BWA, Bowtie2, SNAP…**

! **Mismatches** and **gaps** allowed
→ algorithms have scoring functions

# Taxonomic profiling – recap'

Functional analyses

Marker-based

Relative abundances

Amplicon vs shotgun
Clade-specific markers
(e.g. MetaPhlAn)

Who is there

What about viruses?

# Taxonomic profiling

# Taxonomic profiling – mapping based

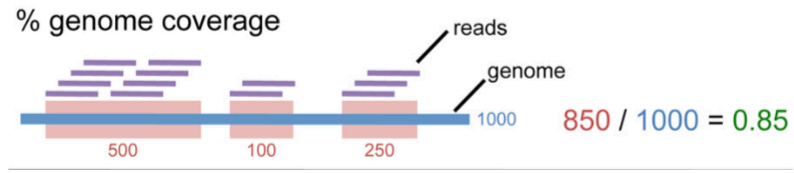
**Label each read with a taxonomy**
(without assembly)


# Taxonomic profiling – mapping based

**Label each read with a taxonomy**
(without assembly)

unknown read

**Reference genomes (RefSeq)**




Virus A
Virus B
Virus C

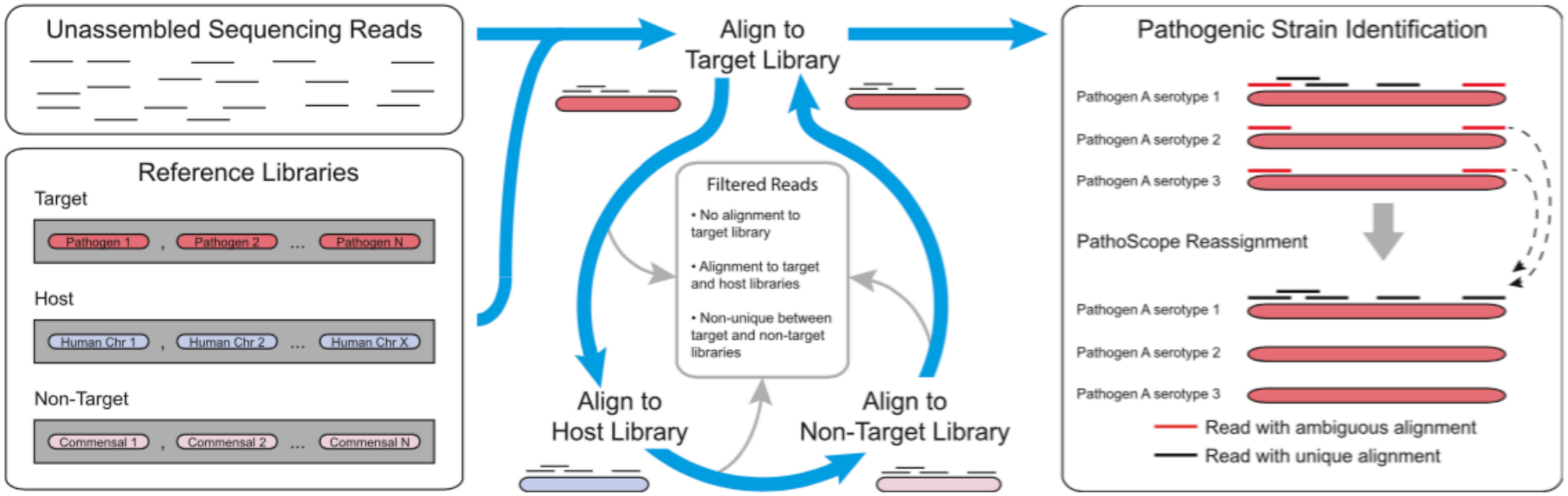

Also works with bacteria

No absolute concentrations

Modified from Petty et al. 2014. J Clin Microbiol. 2014 Sep;52(9):3351-61.

# Taxonomic profiling – mapping based *Examples*

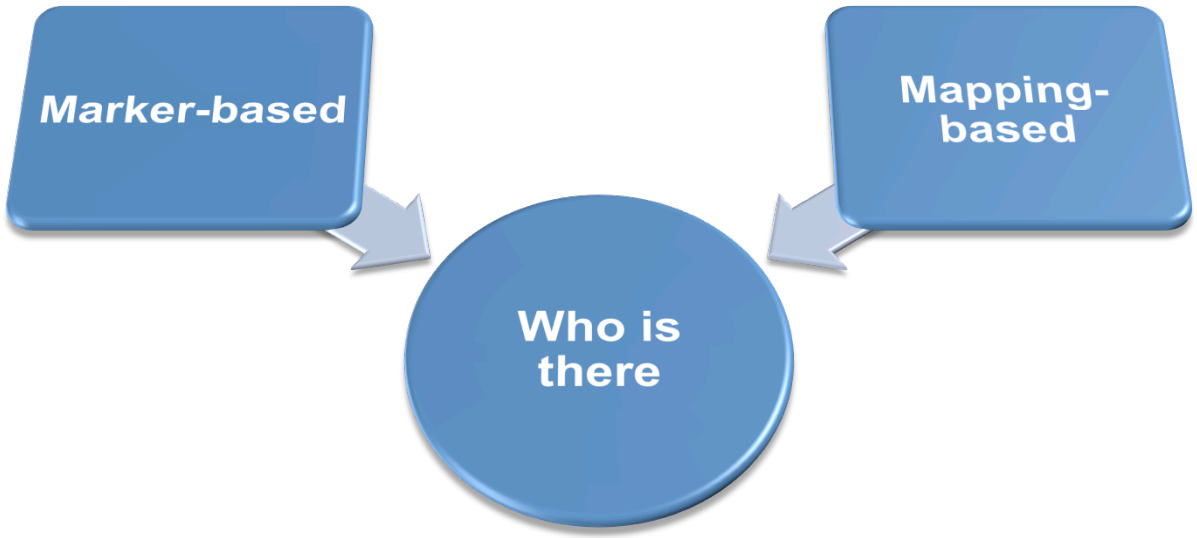
- **ezVIR, SURPI, PathoScope 2.0, Clinical PathoScope**





ezVIR - Petty et al. *J Clin Microbiol*. **52**(9):3351-61 (2014), doi: 10.1128/JCM.01389-14
SURPI - Naccache et al. *Genome Res*. **24**(7):1180-92 (2014), doi: 10.1101/gr.171934.113
PathoScope 2.0 - Hong et al. *Microbiome* **2**:33 (2014) , doi: 10.1186/2049-2618-2-33
Clinical PathoScope - Byrd et al. *BMC Bioinformatics* **15**:262 (2014), doi: 10.1186/1471-2105-15-262 (Image from Byrd et al. 2014)

# Taxonomic profiling – recap'

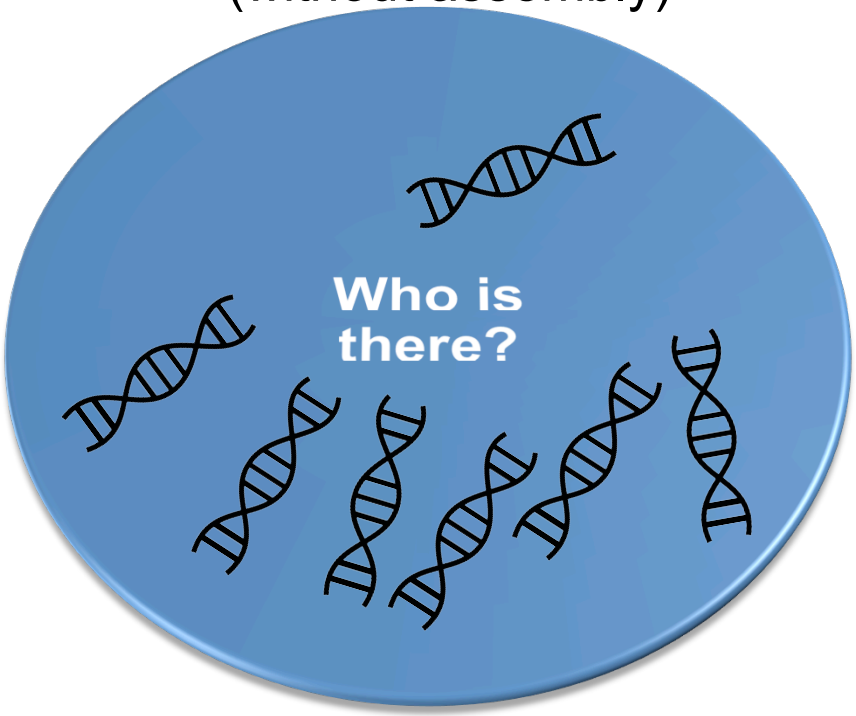



What if we need to be faster?

What other approaches exist?

# Taxonomic profiling – k-mer based

**Label reads using k-mers**
(without assembly)

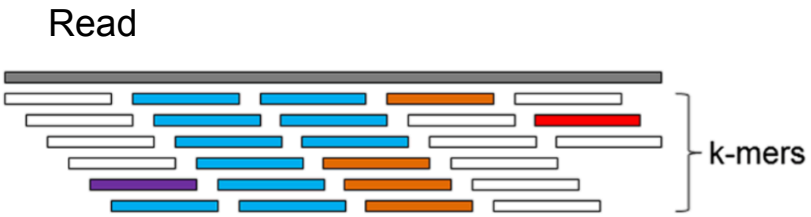


# Taxonomic profiling – k-mer based

## What is a k-mer?

**ATTCGTCATTA…** → List of 5-mers:

ATTCG
TTCGT
TCGTC
CGTCA
GTCAT
TCATT
CATTA




*"Computational sliding of a window"*

## Why does it matter?

✓ Sequence *composition* conservation vs. sequence conservation

Given a list of k-mers for each organism, how to perform the matching of reads?

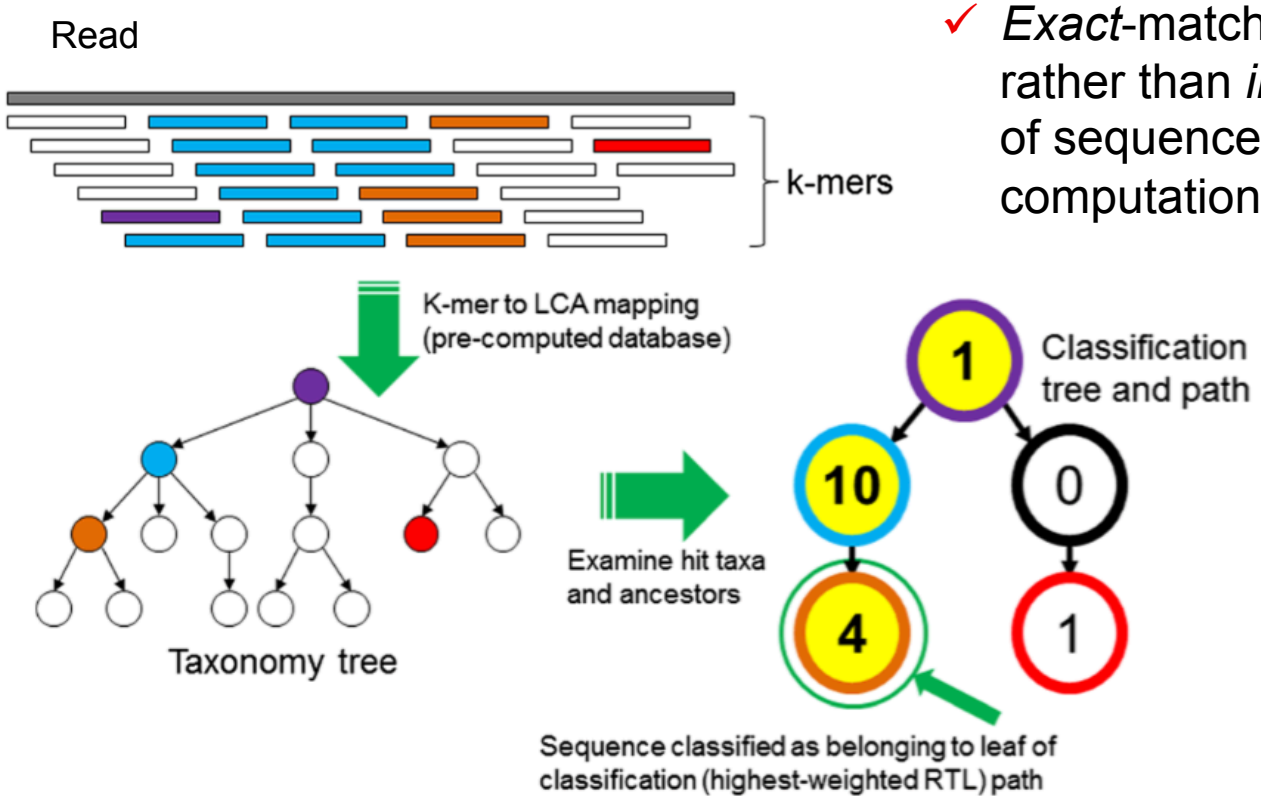
List of k-mers for each organism

*You need some rules…*

Image modified from Wood & Salzburg. *Genome Biol.* **15**(3):R46 (2014), doi: 10.1186/gb-2014-15-3-r46.

# Taxonomic profiling – k-mer based *Examples*

- **Kraken**




✓ *Exact*-match of k-mers, rather than *inexact* alignment of sequences → much easier computational problem

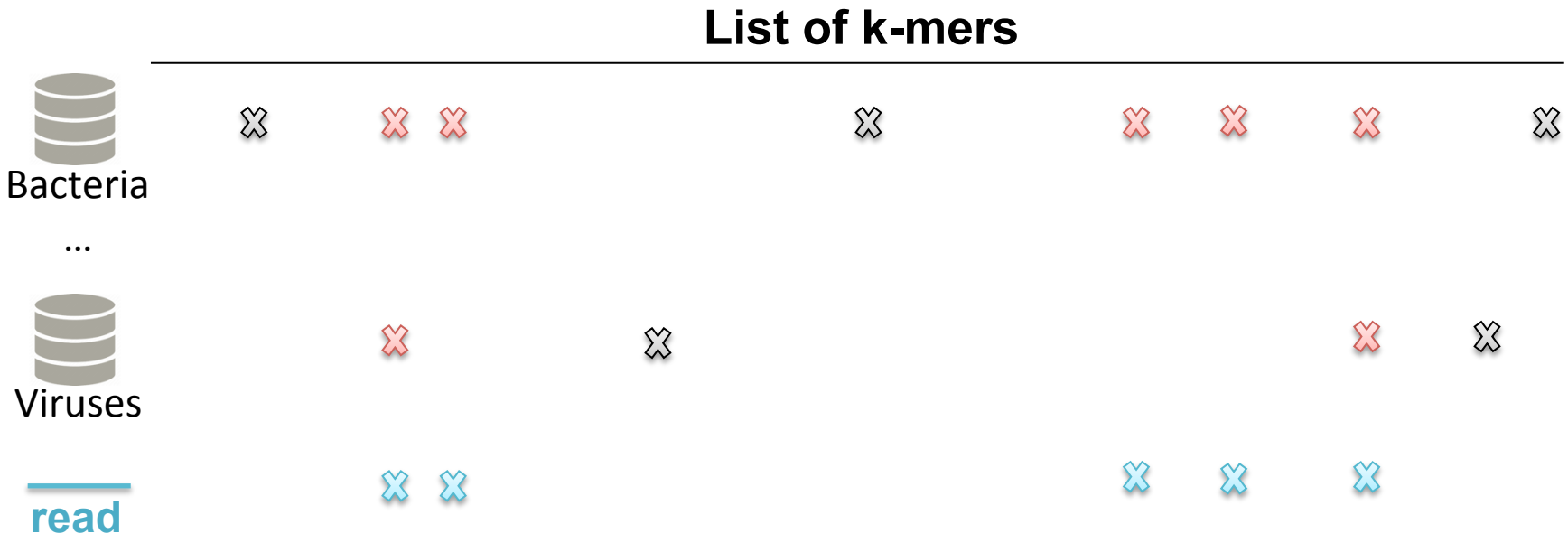
✓ Ultrafast (11x times faster than marker-based MetaPhlAn).
✓ Human subtraction performed during taxonomic profiling

Wood & Salzburg. *Genome Biol.* **15**(3):R46 (2014), doi: 10.1186/gb-2014-15-3-r46.

# Taxonomic profiling – k-mer based *Examples*

- **CLARK, Taxonomer**

BINNING Reads are assigned to the taxonomic group with which most k-mers are shared.

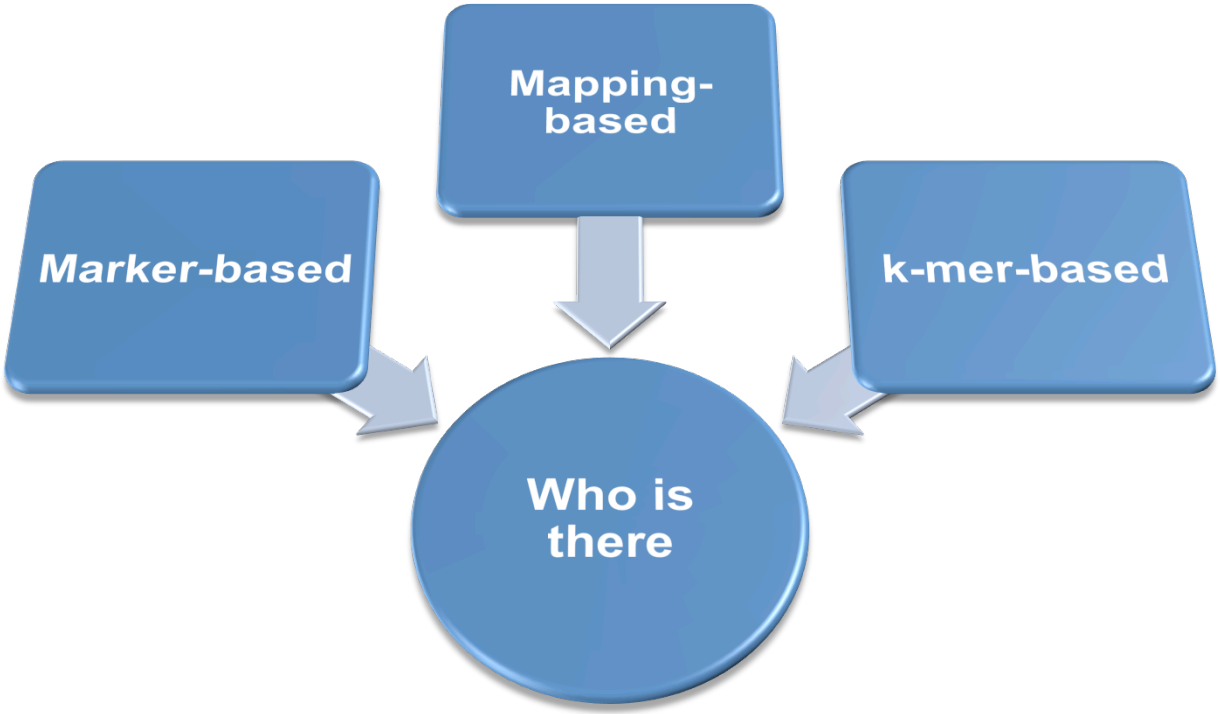



CLASSIF Each read is assigned to the reference that has the maximum total k-mer weight.

- ✓ Almost as fast as Kraken, with more comprehensive taxonomic profiling
- ✓ Also investigates host-expression response profiling (mRNA)

Flygare et al. *Genome Biol.* **17**(1):111 (2016), doi: 10.1186/s13059-016-0969-1

# Taxonomic profiling – recap'

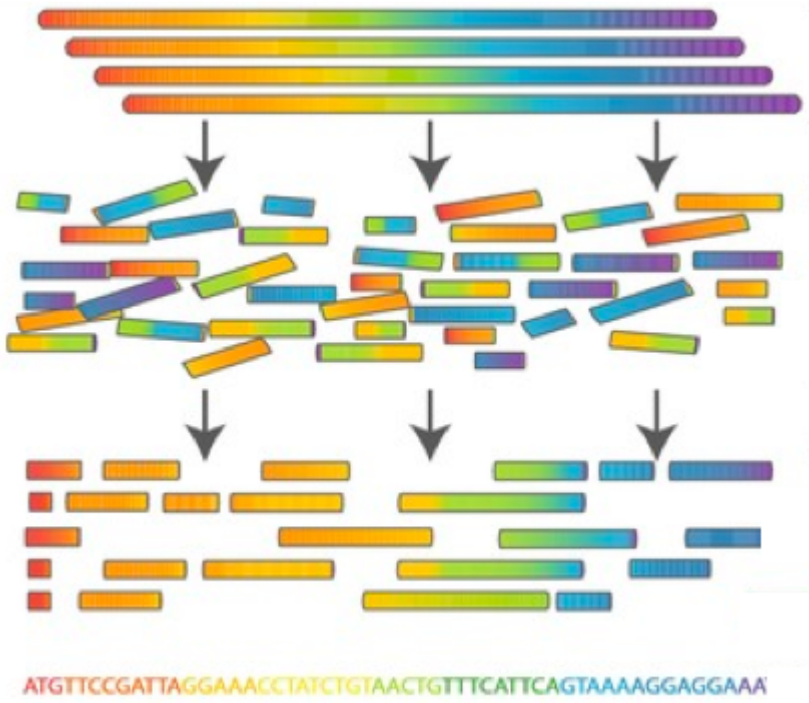



What about unassigned reads?

What about novel pathogens?

# Taxonomic profiling – *de novo* assembly

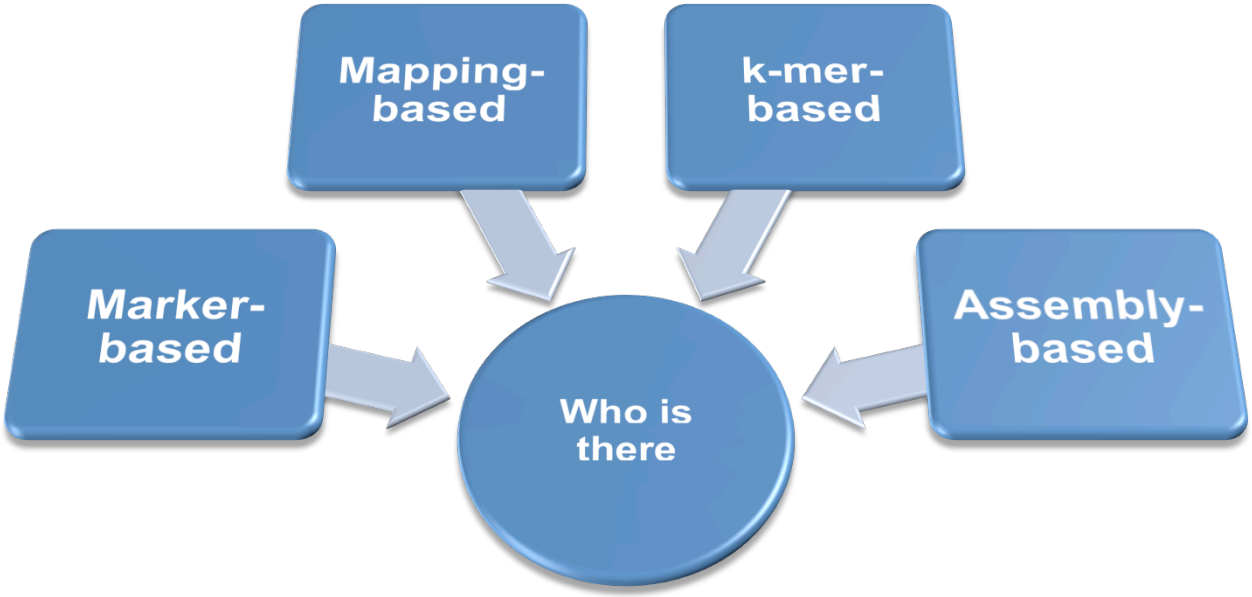
**Assembly first (*de novo)*, then wonder who it is.**
*Usually performed with unlabeled reads ("bag of unknowns").*




✓ (Meta)genome assembly with **de Bruijn** graphs

MetaVelvet - Namiki et al. *Nucleic Acids Res*. **40**(20):e155 (2012)
Meta-IDBA – Peng et al. 2011. *Bioinformatics*. 2011 Jul 1;27(13):i94-101.

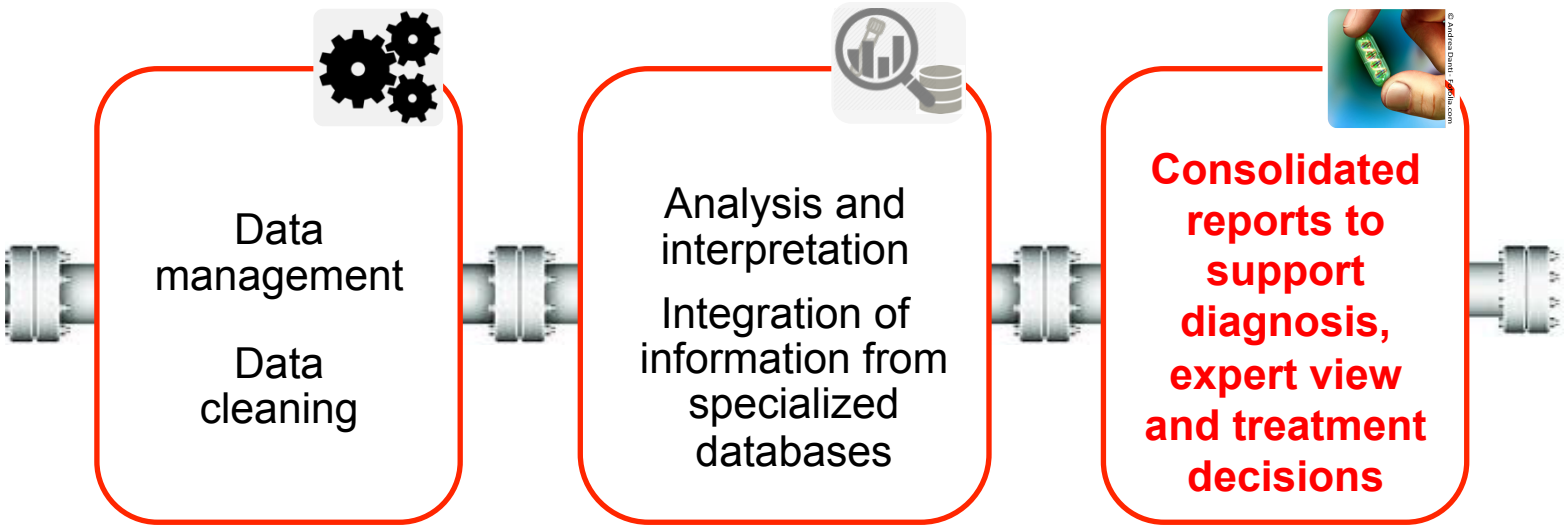
# Taxonomic profiling – recap'




→ **Comparative review:** Peabody et al. *BMC Bioinformatics* (2015) 16:363

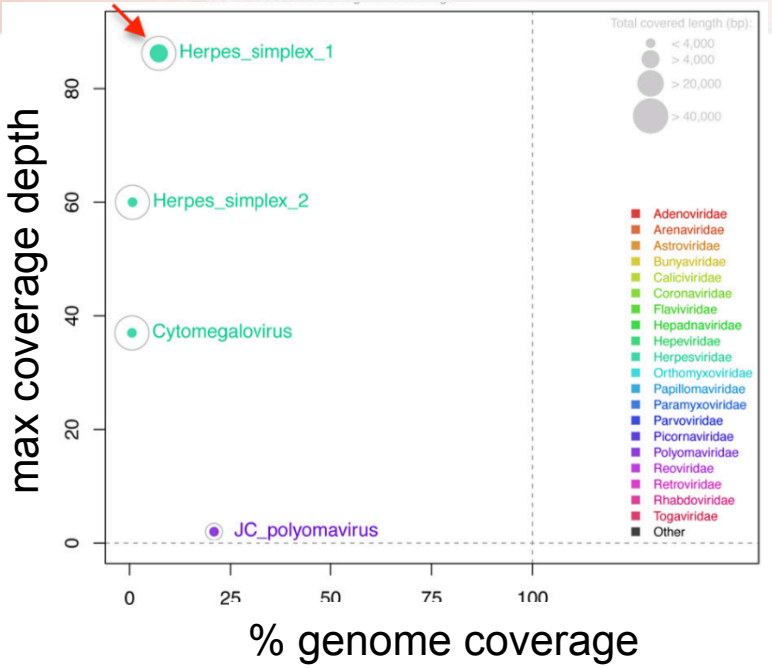
# Report

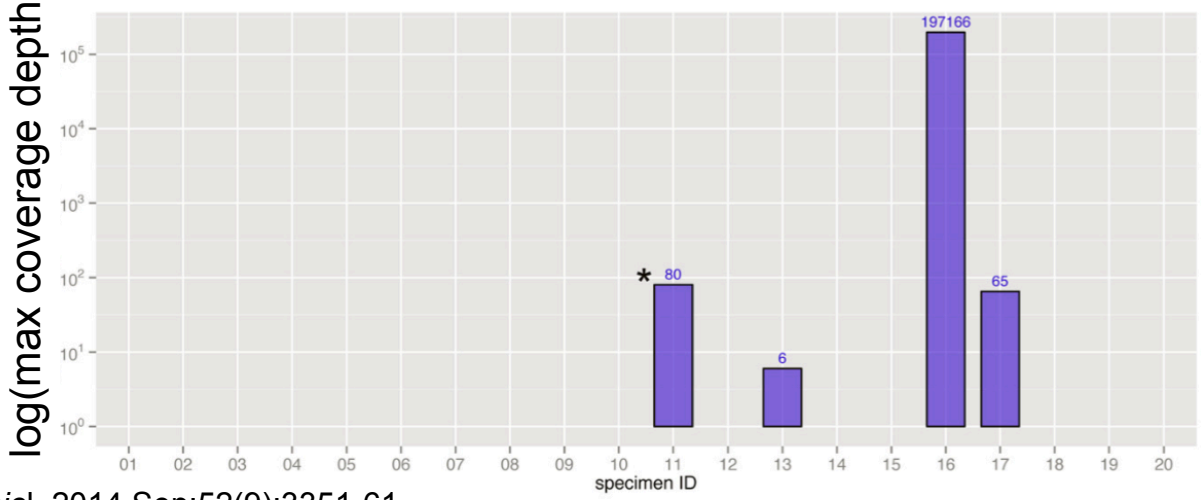
*“Convert metagenomics data into clinically useful knowledge for diagnosis”*



# Report – requirements?

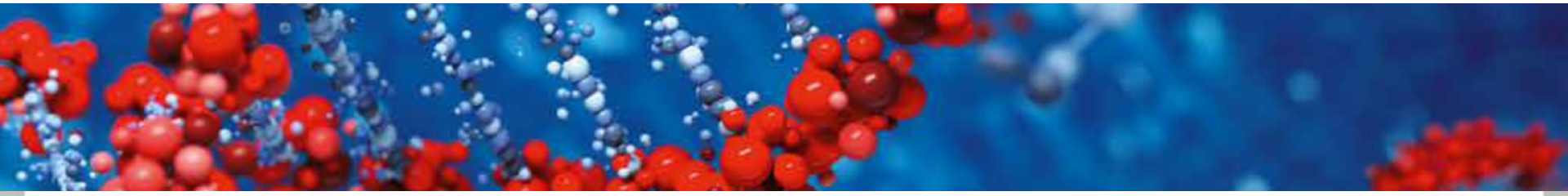
- Quality control
- Spot contaminations during sample preparation
- Clinical interpretation
- Confidence levels
- Absolute concentrations
- Resistance (genotype to phenotype inference)
- Virulence
- …




Modified from Petty et al. 2014. *J Clin Microbiol*. 2014 Sep;52(9):3351-61.

# II. Harmonizing best practices in clinical metagenomics across Switzerland



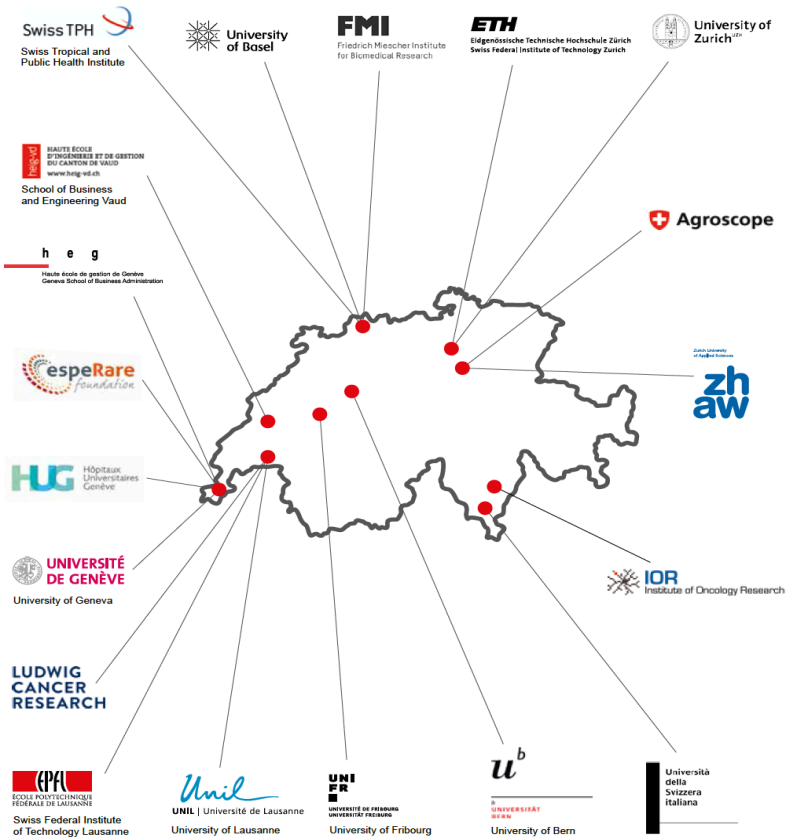
*A Swiss perspective by SIB Clinical Bioinformatics*

# SIB Swiss Institute of Bioinformatics

Swiss Institute of Bioinformatics

## SIB, an efficient collaborative Swiss model

- Swiss-wide institution, federating bioinformatics groups in Switzerland (750 members)
- Leads and coordinates the bioinformatics field in Switzerland
- Recognized leader in bioinformatics (service & infrastructure, research, training)
- Support progress in biological research… and health



# Mission of SIB Clinical Bioinformatics

Provide expertise and support
for the organization, analysis and interpretation of **omics data** for diagnostic purpose, converting them into **clinically-useful knowledge**.

## Trusted partnerships

- Analyze and optimize existing omics pipelines
- Develop, implement and sustain harmonized state-of-the-art tools
- Coordinate involved bioinformaticians (hospitals and SIB)

## Working groups

- Swiss-wide clinical and research stakeholders
- Discussion group to define best practices and harmonize bioinformatics pipelines
- Bridge the gap between research and medical realm

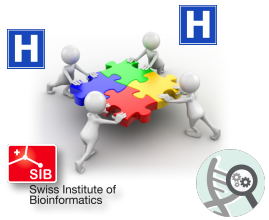
## Training

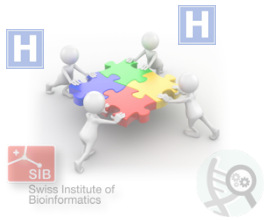
- Provide the required education/training (bioinformaticians, MDs, biologists, students,…)
- Content shaped with clinical stakeholders

# Swiss-wide working groups

## Added values

- ✓ National **consensus**
- ✓ Increased global community **expertise**




**Work Package 1**

Somatic mutations (pathology, hematology)

**Working group 1**

**Work Package 2**

Microbes typing and characterization

**Working group 2**

**Work Package 3**

Metabolomics

**Working group 3**

...

**Work Package N**

Topic N

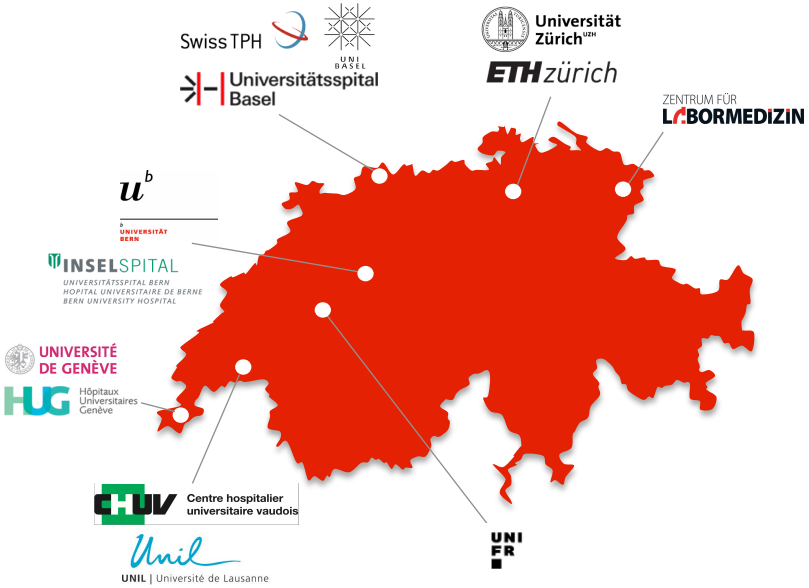
**Working group N**

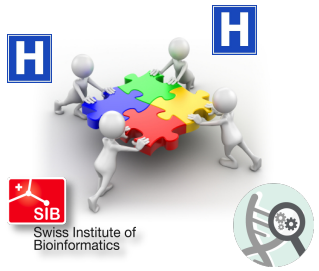
**SIB Steering Committee**

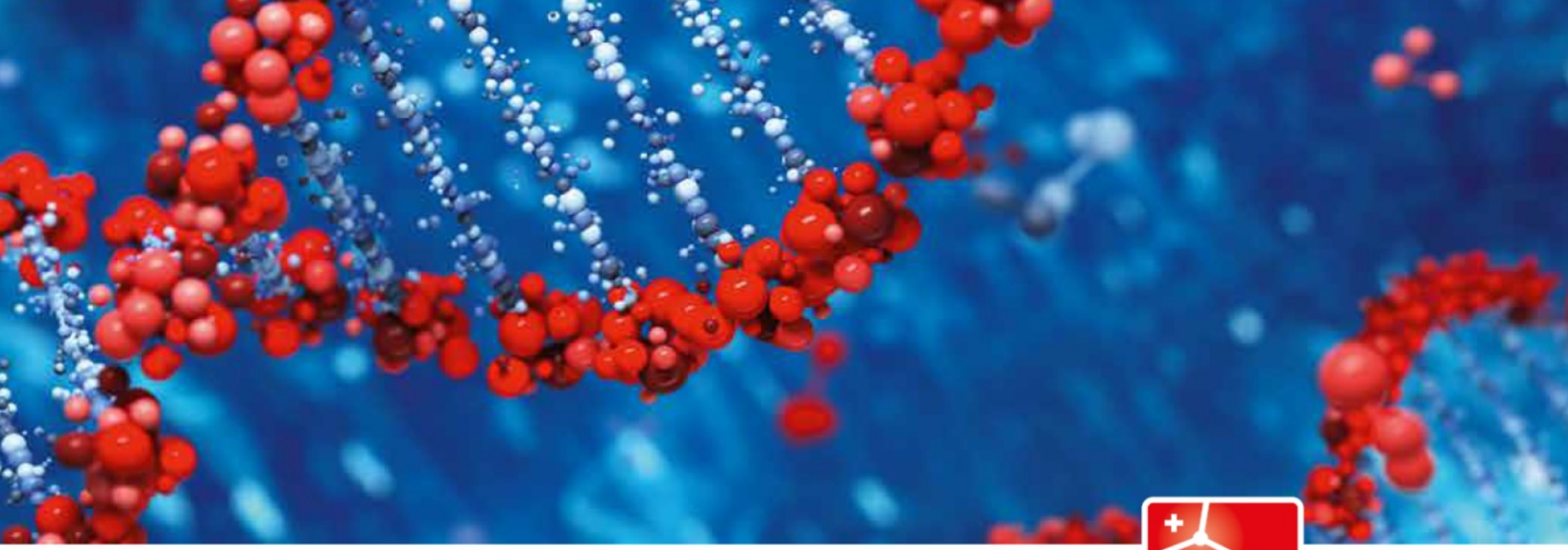
# Swiss-wide working group in microbe typing and characterization (microbiology)

- Co-led by SIB Clinical Bioinformatics (A. Lebrand) and SIB group leader (R. Bruggmann)

- Kick-off workshop in September 2016
  - +50 participants
  - Clinical microbiology labs associated to all university hospitals and research groups
  - Overview of current practices in Switzerland
  - **Identification of main hurdles and needs**
    - ***Define clinical applications***
    - ***Data standardization***
    - ***Benchmarking and harmonization of bioinformatics pipelines***
    - ***Curated reference databases (genomes, genes, proteins)***
    - ***Curated databases for resistance and virulence***
    - ***Training***





Swiss Institute of Bioinformatics

# Thank you for your attention !

## Questions?

*Aitana Lebrand*
*SIB Swiss Institute of Bioinformatics – Clinical Bioinformatics*
***aitana.lebrand@sib.swiss***

elixir SWITZERLAND